Probabilistic Approaches for Communication Avoidance and Resilience in Exascale Simulations

Bert Debusschere and Habib Najm
Sandia National Laboratories, Livermore, CA
{bjdebus, hnnajm}@sandia.gov

Motivation: Among the many fundamental challenges in exascale computing are scalability and resilience. The projected evolutions in communication bandwidth and latency, combined with the dramatic increase in the number of cores, imply that exascale simulations will be much more communication limited than CPU-time limited. Further, the increase in complexity and number of components imply that various system faults, ranging from soft errors such as bit-flips to partial loss of compute nodes, will become much more frequent, to the point that current parallel programming models and approaches for resiliency will be unable to sustain fault-free simulations across many cores for reasonable amounts of time. Key mathematical challenges are a) how to make a simulation less sensitive to communication bottlenecks, b) how to formulate a scientific simulation so it remains well-defined even in the presence of system faults, and c) how to rigorously assess the predictive fidelity of extreme-scale scientific simulations in this context. This white paper offers a path towards addressing these challenges.

We propose that exascale computations should be recast in a probabilistic framework. In this framework, computational results are treated as *data* that is used to update *information* about the solution of the problem, and where this learning problem is cast in a probabilistic setting. In the following, we outline the utility of this framework for communication avoidance, scalability, and resilience.

Probabilistic approaches for communication avoidance: In an exascale computational setting, it is clear that communication costs are dominant. Accordingly, addressing communication costs is paramount to enable efficient utilization of exascale computational platforms. One particularly challenging context where communication is a bottleneck is *all gather* operations. Thus, with N processors, each operating on a subdomain of the problem space, any evaluation of an integral over the computational domain requires gathering contributions from each processor to this integral. For example, in the context of low Mach number reacting flow computations, the poisson equation for the pressure field has a right hand side that involves an integral over the whole domain. Therefore solution of this equation at each time step requires either all-gather/all-scatter or all-to-all operations, as the integral has to be evaluated and be known by each processor.

In this context, the question is, is it possible to avoid waiting for some subset of processors to report their contribution to the sum, and to accept a good approximation of the integral. This is in fact viable in principle when individual processor contributions are considered as data providing information in a probabilistic context where the objective is learning the value of the integral. One can envision a Bayesian learning context, where the value assigned to different subdomains, in terms of their contribution to the integral, can be estimated as the computation progresses, for example, in a time integration context. Accordingly, value judgements can be made, based on the desired accuracy as well any known quantified uncertainties in the computation, regarding the choice of waiting for communication of specific pieces of information versus continuing the computation with the current estimate based on the available information.

Probabilistic reformulation for scalability and resilience: Extending this line of thought to provide scalability and resilience, we propose a novel reformulation of scientific simulations in terms of probability distributions that capture the current state of knowledge about the true solution. Targeted simulations then refine the knowledge about the solution until the distribution converges to the true answer with sufficient

confidence. As such, there is no need to characterize all types of system faults that can occur in a simulation; one focuses solely on the information that a simulation provides, and on using it to reduce the uncertainty in the knowledge about the true solution. As outlined below, the proposed approach is scalable and resilient to faults ranging from soft-errors to loss of compute nodes.

Following this conceptual approach, we developed a probabilistic domain-decomposition preconditioner for the fault-tolerant solution of partial differential equations (PDEs). The algorithm relies on probabilistic representations of subdomain boundary conditions, which, starting from an initial guess, are updated iteratively. In an inner sampling loop, the PDE is solved on each subdomain for sampled values of its current boundary condition distributions. The resulting subdomain solution samples feed into a Bayesian inference of response surfaces that relate the subdomain boundary conditions to each other. Intersection of these response surfaces provides updated samples of the subdomain boundary conditions. When subdomain solves fail (*e.g.* due to nodes crashing), the inference proceeds with fewer samples, and the associated loss of information merely results in a locally higher uncertainty. Erroneous subdomain solves (*e.g.* due to random bit-flips) are either rejected by a properly chosen prior distribution on the solution, or accounted for by the Bayesian noise model. For scalability, the response surface construction and intersection is applied to blocks of subdomains, requiring only local communication, with global updates in an outer sampling loop. With asynchronous communication and distributed data management, this provides scalability with both fine grained concurrency on the level of subdomain solves, and coarser level concurrency on the scale of subdomain blocks. The approach is applicable to both linear and non-linear problems, and is readily compatible with existing solvers as well as instruction-level fault-tolerance approaches.

Preliminary tests using a 1D elliptic model problem show promising convergence results, even under bit-flips. A lot of development work remains, however, in regards to applying this approach to multi-dimensional time-dependent PDEs, or developing a distributed task and data management approach.

Impact: This position paper outlines several approaches: the use of probabilistic estimates for communication avoidance, the reformulation of scientific simulations to allow mathematically rigorous quantification of the effects of a wide variety of system faults, and an iterative approach to solve this reformulated problem in a scalable and resilient fashion. The probabilistic framework made up by these approaches represents a departure from the current deterministic thinking about computer algorithms. This probabilistic approach, however, will be essential to handle challenges caused by communication delays and various system faults in exascale computing, and will help enable such simulations with quantified predictive fidelity.